medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

A case is not a case is not a case - challenges and solutions in determining urolithiasis caseloads using the digital infrastructure of a clinical data warehouse

Martin Schönthaler[1†], Noah Hempen[2*†], Maria Weymann[3], Maximilian Ferry von Bargen[1], Maximilian Glienke[1], Antonia Elsässer[1], Max Behrens[3], Harald Binder[3], Nadine Binder[2]

[1]Department of Urology, Faculty of Medicine and Medical Center - University of Freiburg, Hugstetter Straße 55, Freiburg, 79106, Germany.
[2]Institute of General Practice/Family Medicine, Faculty of Medicine and Medical Center - University of Freiburg, Elsässer Straße 2M, Freiburg, 79110, Germany.
[3]Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center - University of Freiburg, Stefan-Meier-Straße 26, Freiburg, 79104, Freiburg.

*Corresponding author(s). E-mail(s): noah.hempen@uniklinik-freiburg.de;
Contributing authors: martin.schoenthaler@uniklinik-freiburg.de; maria.weymann@uniklinik-freiburg.de; maximilian.vonbargen@uniklinik-freiburg.de; maximilian.glienke@uniklinik-freiburg.de; antonia.elsaesser@uniklinik-freiburg.de; max.behrens@uniklinik-freiburg.de; harald.binder@uniklinik-freiburg.de; nadine.binder@uniklinik-freiburg.de;
†These authors contributed equally to this work.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Abstract

Background:
To provide more evidence in urolithiasis research, we have established the German Nationwide Register for RECurrent URolithiasis (RECUR) using local clinical data warehouses (CDWH). For RECUR and other registers relying on digitalized clinical data, it is crucial to ensure the data's reliability for answering scientific questions. In this work, we aim to compare the results of different CDWH-based queries on urolithiasis cases next to manual case extraction from the primary source.
Methods: Sources for data extraction included the Medical Center University of Freiburg (MCUF) hospital information system (HIS), MCUF performance data (a clinical data set with merged data from patients including data from various time points throughout their treatment), and MCUF reimbursement data. We extracted data on caseloads in urolithiasis algorithmically (performance and reimbursement data) and compared those to a reference group compiled of manually extracted data from the local HIS and algorithmically extracted data.
Results: Algorithmic extraction based on performance data resulted in correct and complete case identification as compared to the reference group. The case numbers from manual extraction from HIS data and algorithmic extraction from reimbursement data differed by 14% and 12%, respectively. The reasons for deviations in HIS data included human errors and a lack of data availability from different wards. Deviations in reimbursement data arose primarily due to the merging of cases in the context of reimbursement mechanisms. As the CDWH at MCUF is part of the German Medical Informatics Initiative (MII), the results can be transferred to other medical centers with similar CDWH structure.
Conclusions: The current study provides firm evidence of the importance of clearly defining a study's target variable, e.g., urolithiasis cases, and a thorough understanding of the data sources and modes used to extract the target data. Our work clearly shows that, depending on various data sources, a case is not a case is not a case.

Keywords: clinical routine data, medical informatics, data integration, secondary use of clinical data, real-world data, clinical registry, data integration centers, Medical Informatics Initiative, data mining, urolithiasis

1 Background

Urolithiasis is a widespread disease that has a tremendous impact on both individuals and societies worldwide. Patients suffer from recurring episodes of intense pain that require outpatient or inpatient treatments, including surgical interventions for stone removal. Long-term morbidity includes chronic kidney disease and arterial hypertension [1]. However, the level of evidence in urolithiasis research tends to be low [2]. On

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

the other hand, we see an increasing amount of routine data collected systematically [3–11]. Authors propose its use for observational studies, registers, and other electronic data sets, to fill critical gaps in evidence [12]. This motivated the authors to propose a German nationwide registry for RECurrent URolithiasis in the Upper Urinary Tract (RECUR) in 2017 [1], which was funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF). We aim to establish this fully automated register from routine data collected in electronic health-care systems, in our case, relying on the digital infrastructure of the German Medical Informatics Initiative (MII). The MII has been launched to create a digital network connecting all German medical university hospitals, based on local clinical data warehouses (CDWH). [13, 14]

CDWHs host integrated, standardized, and pseudonymized routine clinical data from heterogeneous primary sources. [3, 15–19] As designed for scientific co-use, some key challenges they address are data harmonization, standardization, and data quality issues. [15, 20] These challenges are reflected in the 4 phases of data flow in CDWH described by Doutreligne et al.: data collection, transformation, provision, and usage. [20] Other authors describe three other aspects regarding data processing in CDWH: data integration, consolidation, and presentation. [21–23] The MII intends to create internationally compatible data. Therefore, all participating hospitals are setting up CDWHs by incorporating standardized data sets including basic patient data, laboratory results, or medication. CDWHs are set up and operated by local data integration centers (DIC). The subsequent data sets are transferred into research data repositories of the local DIC. As the DIC Freiburg is part of the MII, the accessible data have a similar structure and characteristics to those of other medical centers in Germany. [13]

In the second step, case registers can be derived using CDWHs and routine data. Two German examples are the German Pain e-Registry [24] and the German Chest Pain Registry, where the latter was set up from manually extracted routine data. [25] The

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

manual extraction of clinical data is very time-consuming. Depending on the disease's prevalence, it quickly becomes unmanageable and relies on trained personnel who are required to minimize human error. However, there is already an area in which trained personnel have been systematically recording routine clinical data for several years, namely, in the billing of hospital cases.

With the introduction of diagnosis-related groups (DRGs) in the German reimbursement system, a different research focus on routine clinical data has emerged, concentrating on using routine data generated primarily for reimbursement purposes. German hospitals are obligated to transfer DRG data to the German DRG Institute (Institut für das Entgeltsystem im Krankenhaus GmbH – INEK). [26] Since 2005, the German Federal Statistical Office (Deutsches Statistisches Bundesamt - DESTATIS) has provided statistics from official data on several areas, including society, the economy, the environment, and the state. DESTATIS also enables direct access to the data for further analysis, including the DRG data provided by hospitals. Several medical-scientific studies using these DESTATIS data have already been published. [27–32] Specifically, a recently published study evaluated trends in the incidence of urolithiasis and the use of therapeutic interventions in Germany between 2005 and 2016. [33] However, reimbursement data are subject to certain coding requirements and conditions, e.g., two inpatient stays of one patient that occur within 30 days and that are clinically related typically count as a single case in the DRG system. In this case, two clinically distinct cases are counted as one case in the DRG data, which then leads to an incorrect count of actual cases.

Concerning RECUR and potential initiatives for building repositories and registries based on digitalized routine clinical data, it is essential to ensure that the requested data are reliable for answering scientific questions. Investigators should identify cases correctly according to their case definition.

In this work, we aimed to compare and validate the results of different CDWH-based

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

queries for urolithiasis cases next to manual case extraction from the primary source. We take a closer look at the structure of routine clinical data in the context of a CDWH by illustrating the clinical data flow from the primary clinical data source to reimbursement data of the Medical Center - University of Freiburg (MCUF) and the differences in the resulting data sources. We specifically aimed to study two aspects: first, to determine the complete number of urolithiasis cases treated at the Department of Urology of the MCUF during a predefined 2-month period; second, to identfy a reliable data source for the correct identification of all relevant urolithiasis cases. To the best of our knowledge, this is the first study to systematically investigate differences between various data sources and extraction modes in a urological context. However, this approach is essential for providing reliable results when using routine clinical data derived from CDWH for scientific use.

2 Methods

2.1 Study population and case definition

We included patients older than 18 years with urinary stones in the upper urinary tract admitted for in-hospital treatment at the Department of Urology of the MCUF between December 1st, 2020, and January 31st, 2021. To be considered an inpatient case, patients had to stay overnight. We defined a case as any patient treated for urolithiasis as encoded in the local hospital information system (HIS) using the International Classification of Diseases (ICD Version 10 German Modification = ICD-10-GM; October 1st, 2015; Table 1). A primary diagnosis is the main diagnosis that leads to hospitalization. The secondary diagnosis represents comorbidities that were relevant during the hospital stay, e.g., arterial hypertension requiring treatment. Our case definition included both patients with a primary urolithiasis diagnosis and patients with a primary diagnosis of urolithiasis-associated complications (e.g., hydronephrosis, renal colic, or urinary tract infection) combined with a secondary

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

diagnosis of urolithiasis (Table 1). Patients readmitted after midnight of the day of discharge were considered separate cases.

2.2 Clinical Data Flow and Extraction of Data

The sources for data extraction were (1) the HIS of the MCUF, (2) the MCUF performance data, and (3) the MCUF reimbursement data. Figure 1 illustrates the clinical data flow between these three sources, the corresponding integration into the DIC, and the consecutive data extraction. We extracted data algorithmically (performance and reimbursement data) and manually (HIS) for validation. In the following, we briefly describe the three different data sources and extraction modes.

Mode 1: Manual extraction from hospital information system

The primary source for all data used in this study was the local HIS implemented in an application called Prometheus. This comprehensive, integrated information system is designed to support all aspects of patient management. Prometheus integrates all personal and clinical patient data, such as ICD and OPS (Operation and Procedure Classification System) encodings, laboratory and imaging results, and medical reports. One coauthor (M. F. B.) performed the manual data extraction from the Urology section of Prometheus. The medical records of all patients admitted to the Department of Urology during the designated period were screened. Inpatient stay-related information on all patients with documented relevant ICD codes (table 1) was extracted. All digital records (medical and surgical reports, laboratory and radiology reports, etc.) of these patients were checked regarding the case definition. A consultant urologist (M.S.) verified the final list of manually extracted cases and all related information, including primary and secondary diagnoses, major and minor diagnosis status, dates of hospital admission and discharge, age of patient, and treatment.

Mode 2: Algorithmic extraction from performance database

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

In the second step of data processing at the MCUF, trained personnel transfer and integrate ICD and OPS codes from various HIS sources (including Prometheus) into a separate internal performance database. This database was set up as an intermediate step for deriving reimbursement data. The ICD codes are generated at multiple stages of patient management, e.g., on admission, at discharge, at patient transfer to a different ward or department, or during treatment. Performance data on included patients were digitally transferred into a clinical data repository (CDR) of the DIC in a standardized format prespecified by all university hospitals participating in the MII throughout Germany. Based on the inclusion criteria, an algorithmic query was developed and run on the performance database to extract the urolithiasis cases and related case information.

Mode 3: Algorithmic extraction from reimbursement database

In the third step, trained personnel at the MCUF generate reimbursement data from the performance data following the DRG guidelines. This step is mandatory for all hospitals in Germany, which have created digitally available reimbursement data since 2003. The MII initially integrated these data into the respective DIC's scientific data repository (SDR). The data were depicted in a standardized format as part of the so-called MII core data set to enable cross-site scientific evaluations. This processing also revealed that distinct inpatient stays of patients treated two or more times within 30 days could have been merged into a single case for reimbursement, following DRG guidelines. In the same way as for mode 2, the algorithmic query was performed according to the inclusion criteria on the SDR to extract urolithiasis cases and associated case information from the reimbursement data.

2.3 Outcome measures

Based on the case definition described above, we compared the number of cases extracted using the three modes. To assign corresponding cases from manual and

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

algorithmic extractions, pseudonymized data from the CDR and SDR were decoded. In addition, we compared the dates of admission and discharge of the included patients, primary and secondary ICD codes at hospital discharge, information on the involvement of the Department of Urology (in patients admitted to other departments), and the specific wards of the MCUF where a patient was treated. Any discrepant cases were discussed among the team of physicians and trained encoding personnel. We compiled a reference group of all cases found by manual extraction and added missing cases from algorithmic extraction. In cases of discrepancies between extracted cases using any of the modes described above and the reference group, we explored potential causes.

3 Results

We identified 47 individual patients treated for urolithiasis in the upper urinary tract at the MCUF from December 1st, 2020, through January 31st, 2021. Of those, 11 individuals were admitted twice during that period, resulting in two cases per individual. A total of 58 cases were included, constituting the reference group. A comparison of all three extraction modes revealed the following: manual extraction from HIS data (mode 1) correctly identified 52 cases. Algorithmic case extraction from performance data (mode 2) correctly identified all 58 cases, while algorithmic case extraction from reimbursement data (mode 3) correctly identified 52 cases. Table 2 provides a detailed breakdown of the number of cases identified using the extraction modes as well as various discrepancies. The reasons for discrepancies in the number of cases identified, i.e., missing cases, were as follows:

First, of the six missing cases from the HIS data extraction (mode 1), three were missed during manual extraction by the extractor but could be subsequently verified;

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

the other three cases were urology patients who were admitted to a non-urology department due to a short-term bed shortage. These cases were recorded in the respective ward section in Prometheus, which was not screened by the extractor. The three cases could also be subsequently verified. During the first run of the manual HIS data extraction, we also found two additional cases that were later excluded after we found a discrepancy with our case definition.

Second, algorithmic extraction of the reimbursement data (mode 3) missed 6 cases, four of which were combined in the reimbursement data (patients treated twice or more often within 30 days merged into a single case for reimbursement). In addition, we identified one case with a primary urolithiasis diagnosis and a non-urolithiasis secondary diagnosis. This was also a case that was merged based on the DRG guidelines in which the more complex treatment was then coded via the major diagnosis, in this case, septicemia.

In addition to missing cases, we found cases involving incomplete or incorrect documentation, specifically concerning data on admittance and discharge dates, ICD codes at discharge, and the involvement of the Department of Urology. HIS data extraction revealed three cases recorded in Prometheus with an incorrect admission date, one case with an incorrect discharge date, and three cases with missing documentation from the ward where patients were treated. We also identified 23 cases showing minor discrepancies in diagnosis coding. However, none of these discrepancies affected the recognition as a case. The algorithmic extraction of the performance data revealed one case with an incorrect discharge date and seven cases with minor discrepancies in the subclassifications of the ICD codes, again not affecting their assignment as urolithiasis cases. The algorithmic extraction of reimbursement data revealed three cases in which the primary urolithiasis diagnosis and a secondary diagnosis of an

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

associated complication had been interchanged, thus still fulfilling the case definition. We also found eight cases with incorrect discharge dates, five thereof attributable to merging cases for reimbursement. Three cases showed minor discrepancies in ICD code subclassifications. Neither of these discrepancies affected the assignment of those patients.

4 Discussion

RECUR is built upon the digital infrastructure of the MII, which was launched to create a digital network connecting all German medical universities based on CDWHs. The CDWH scientific data repositories collect extracted data from a local HIS. Within an HIS, patient data sets may vary depending on the mode of operation. At the MCUF, we identified three potential sources for data extraction—modes 1 to 3—as described above. In this study, we investigated caseloads in urolithiasis as a use case to analyze different data sources in a CDWH. We searched for potential discrepancies when extracting data and for an extraction mode that best reflects inpatients.

This investigation and the resulting findings constitute a crucial step toward enhancing the validity of RECUR or any other register based on data from CDWH as established within the MII. We outlined the defined case criteria and obtained the required information from data extracted manually and algorithmically from the CDWH of the MCUF. We established a reference group of urolithiasis patients to compare the different data sources. We have demonstrated that algorithmic extraction based on "performance data" – a clinical data set constituting merged data from "cases" of an individual treated at different departments – resulted in correct and complete case numbers based on our study's case definition. In contrast, 14% and 12% of the cases derived from manual extraction of the primary HIS data and algorithmic extraction of the reimbursement data, respectively, deviated from the

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

reference group. We identified two main reasons for this: (i) the merging of cases following the coding guidelines in the German DRG reimbursement system and (ii) human, i.e., cases overlooked in manual extraction.

To use routine data reliably, it is essential to understand their structure. To our knowledge, this is the first study to systematically examine the differences between different sources of routine clinical data and extraction methods in a urologic context. A systematic investigation of extraction modalities was carried out during a specific period to identify all urolithiasis cases treated inpatient at the Department of Urology at the MCUF. Three researchers independently validated the extracted data, which enabled us to identify missing information in the data extractions and form a complete reference group. Based on the results of this study, we will identify the correct data sources to obtain valid data for RECUR. From a broader perspective, the findings of this study may be applied to all scientific data analysis based on clinical data from hospital information systems.

Our study has some limitations. We investigated a rather short period of two months in 2020/2021, resulting in a rather small sample of relevant cases. However, we have identified that this period is sufficient to detect relevant differences between data sources and extraction modes for answering the question of interest. Furthermore, the practice of data entry and further processing, including coding into reimbursement data in the CDWH, has been stable since then, so the results can be directly translated to today's standards. We consider data from the CDWH of the University of Freiburg, which may have its standards for data processing, including a specific hospital information system and processing into the performance database. However, the reimbursement structures and coding guidelines for reimbursement data are standardized across Germany. [26, 34] ICD codes used for registries should clearly define

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

the clinical conditions of patients. To enhance accuracy in case identification, which can be deficient when using only ICD codes [35], additional OPS codes (encodings for treatments) can be used. In our study, however, ICD codes were sufficient. As a general limitation of any data extraction, we also had to rely on the completeness and correctness of the data sources used.

Based on our results, we emphasize the following aspects to achieve reliable results from data extraction using a CDWH:

1. Clear definition of the target variable (in our study, *urolithiasis cases*)
2. Identification of the corresponding variables in available clinical data sets (in our study, *ICD codes encoding urolithiasis*)
3. Identification of the appropriate clinical data source to provide comprehensive data on the target variable (in our study, *performance data*)

For RECUR, we conclude that the MCUF performance data set is an appropriate database. Our local CDWH/DIC will extract all the data for the RECUR scientific data repository from the MCUF performance data. To ensure the validity of the RECUR data from other participating hospitals, the data sources must be investigated and validated in the same way. In general, this work shows how important it is to understand the structure of routine data to be later used for analytical purposes.

5 Conclusion

RECUR, as an automated digital register, will depend on the availability and reliability of scientific data repositories, as established by the data integration centers of the German Medical Information Initiative. This requires gradual adaptation or better yet standardization of formal and technical approaches in both the HISs of participating hospitals and the CDWHs of their respective DICs. In general, the current study provides firm evidence of the importance of clearly defining the target variable of a

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

study, e.g., urolithiasis cases, and a thorough understanding of the data sources and modes used to extract the target data. This work has clearly shown that a case is not a case is not a case.

6 List of abbreviations

RECUR - Register for RECurrent URolithiasis

BMBF - German Federal Ministry of Education and Research

MII - German Medical Informatics Initiative

MCUF - Medical Center University of Freiburg

DIC - Data Integration Centre

CDWH - Clinical Data Warehouse

HIS - Hospital Information System

CDR - Clinical Data Repository

SDR - Scientific Data Repository

ICD - International Classification of Diseases

OPS - Operation and Procedure Classification System

DRG - Diagnosis Related Groups

INEK - Institut für das Entgeltsystem im Krankenhaus GmbH, German DRG Insitute

DESTATIS - Deutsches Statistisches Bundesamt, German Federal Statistical Office

MIRACUM - Medical Informatics in Research and Care in University Medicine

DFG - Deutsche Forschungsgemeinschaft, German Research Foundation

Declarations

Ethics approval and consent to participate. The protocol for the collection and management of data was approved by Ethics Committee, Medical Center – University of Freiburg, Freiburg, Germany (Reference number 40/20) and carried out according

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

to the Declaration of Helsinki. The lack of need for informed consent was confirmed by the ethics committee named previously. Additional approval for data use was obtained by the Use & Access Committee (UAC) of the Medical Center – University of Freiburg.

Consent for publication. Not applicable

Availability of data and materials. The data that support the findings of this study are not openly available due to reasons of sensitivity.

Availability of code. Not applicable

Competing interests. The authors declare that they have no competing interests.

Funding. RECUR is funded by the German Federal Ministry of Education and Research (BMBF, FKZ 01GY1902). MIRACUM is funded by the German Federal Ministry of Education and Research (BMBF) within the "Medical Informatics Funding Scheme" (FKZ 01ZZ1606H). The work of N. Hempen, M. Behrens, H. Binder and N. Binder has been funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 499552394 – SFB 1597. The work was supported by the Department of Urology, the Institute of General Practice/Family Medicine, and the Institute of Medical Biometry and Statistics, Medical Center – University of Freiburg, Freiburg, Germany.

Authors' contributions. M.S., N.B., H.B., and M.W. conceived and designed the study. M.F.B., M.G., N.B., and M.B. completed data extraction. N.H. performed data analysis and provided tables and figures. N.H., M.S., and N.B. wrote the manuscript. M.W. and A.E. provided technical guidance and revised the manuscript. All authors approved the final manuscript.

Corresponding Author. Correspondence to Noah Hempen.

Acknowledgements. Not applicable

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Figures

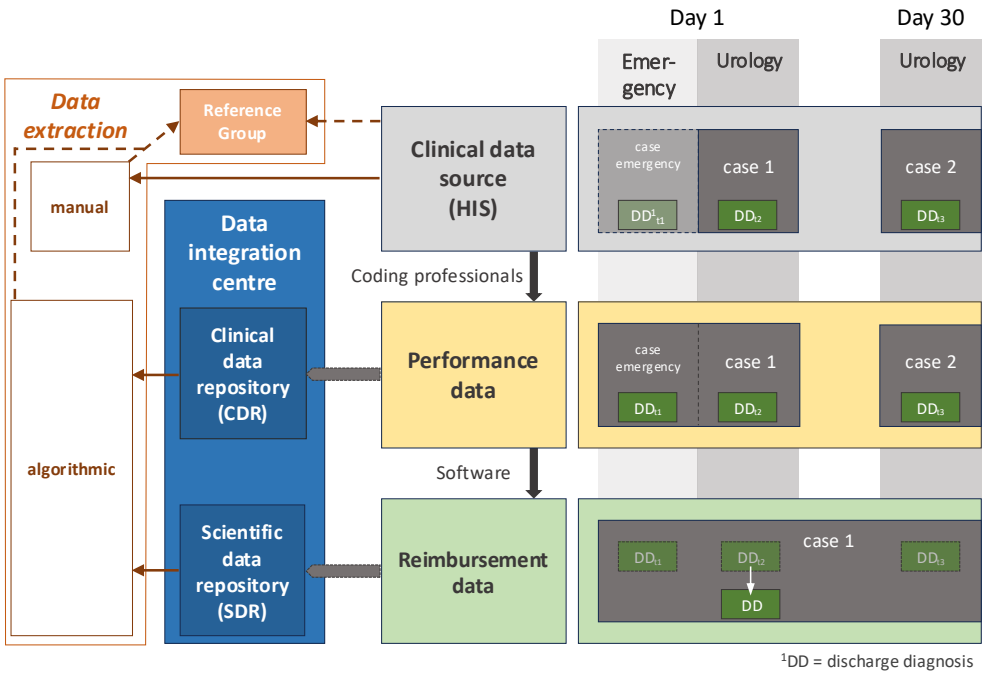

Fig. 1 Clinical data flow and data extraction modes. (1) The Hospital Information System houses all personal and clinical patient data using various software/clinical management systems: separate "cases" for an individual patient treated by various departments and/or timings of treatments (gray box). (2) Performance data are extracted by trained coding professionals: relevant data from "cases" of an individual treated at different departments are merged (yellow box). The performance data are integrated into the clinical data repository (CDR) of the DIC. (3) Reimbursement data: "Cases" of an individual are merged within 30 days, which would make one case out of actually two cases (green box). The reimbursement data were initially integrated into the scientific data repository (SDR) of the DIC. From all three data extraction modes (left box), the reference group (orange box) of all correctly identified urolithiasis cases was constructed after careful validation.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Tables

Table 1 ICD codes of relevant diagnoses.

Primary diagnosis (ICD-10)		Secondary diagnosis (ICD-10)
Urolithiasis codes		
N13.2	Hydronephrosis with renal and ureteral calculous obstruction	Any secondary diagnosis
N20.0	Calculus of kidney	
N20.1	Calculus of ureter	
N20.2	Calculus of kidney with calculus of ureter	
N20.9	Urinary calculus, unspecified	
N22.0	Urinary calculus in schistosomiasis	
N22.8	Calculus of urinary tract in other diseases classified elsewhere	
Associated codes		
N13.1, N13.3 to N13.9	Hydronephrosis, pyonephrosis, etc.	Urolithiasis codes
N23	Unspecified renal colic	
N39.0	Urinary tract infection	
A41	Sepsis	

ICD codes of the relevant diagnoses representing the inclusion criteria for the study population – i.e., identifying all urolithiasis cases. The primary diagnosis is the main diagnosis that led to the patient's hospitalization. The secondary diagnosis represents comorbidities that were relevant during the hospital stay. In our study, we included patients with a primary diagnosis of urolithiasis and patients with a primary diagnosis of urolithiasis-associated complications (as combined with a secondary diagnosis of urolithiasis).

Source: ICD Version 10 German Modification = ICD-10-GM; October 1st, 2015

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Table 2 Identified cases.

	Case number Reference group	HIS data Mode 1 - *manual extraction*	Performance data Mode 2 - *algorithmic extraction*	Reimbursement data Mode 3 - *algorithmic extraction*
Correctly identified cases	**58**	**52**	**58**	**52**
Missing cases				
Merging of cases[1]	-	-	-	4
Change of primary and secondary diagnosis[1]	-	-	-	1
Urolithiasis patient admitted to another department[2]	-	3	0	1
Human errors	-	3	0	0
Additional cases[3]				
Merging of cases[1]	-	-	-	1
Human errors	-	2	-	-

Comparison of correctly identified urolithiasis cases from the three different extraction modes. Reasons for missed cases include the merging of cases, changes in diagnoses, patients not being listed in a ward run by the Urology department as well as human errors. One additional case was identified incorrectly due to the merging of cases for reimbursement purposes.

[1]due to reimbursement issues

[2]other than Urology (due to shortage of beds in Urology)

[3]identified correctly

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

References

[1] Schoenthaler M, Fichtner UA, Boeker M, Zoeller D, Binder H, Prokosch HU, et al. A nationwide registry for recurrent urolithiasis in the upper urinary tract – The RECUR study protocol. BMC Health Services Research. 2022 Aug;22(1):1060. https://doi.org/10.1186/s12913-022-08375-7.

[2] Schoenthaler M, Miernik A, Wilhelm K, Schlager D, Schoeb DS, Adams F, et al. Level of evidence, sponsorship, conflict of interest policy and commercial impact of PubMed-listed clinical urolithiasis-related trials in 2014. BJU International. 2016 May;117(5):787–792. https://doi.org/10.1111/bju.13387.

[3] Jha AK, DesRoches CM, Campbell EG, Donelan K, Rao SR, Ferris TG, et al. Use of Electronic Health Records in U.S. Hospitals. New England Journal of Medicine. 2009 Apr;360(16):1628–1638. https://doi.org/10.1056/NEJMsa0900592.

[4] DesRoches CM, Charles D, Furukawa MF, Joshi MS, Kralovec P, Mostashari F, et al. Adoption Of Electronic Health Records Grows Rapidly, But Fewer Than Half Of US Hospitals Had At Least A Basic System In 2012. Health Affairs. 2013 Aug;32(8):1478–1485. https://doi.org/10.1377/hlthaff.2013.0308.

[5] Kanakubo T, Kharrazi H. Comparing the Trends of Electronic Health Record Adoption Among Hospitals of the United States and Japan. Journal of Medical Systems. 2019 Jul;43(7):224. https://doi.org/10.1007/s10916-019-1361-y.

[6] Sheikh A, Jha A, Cresswell K, Greaves F, Bates DW. Adoption of electronic health records in UK hospitals: lessons from the USA. The Lancet. 2014 Jul;384(9937):8–9. https://doi.org/10.1016/S0140-6736(14)61099-0.

[7] Apathy NC, Holmgren AJ, Adler-Milstein J. A decade post-HITECH: Critical access hospitals have electronic health records but struggle to keep up with other

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

advanced functions. Journal of the American Medical Informatics Association. 2021 Aug;28(9):1947–1954. https://doi.org/10.1093/jamia/ocab102.

[8] Liang J, Li Y, Zhang Z, Shen D, Xu J, Zheng X, et al. Adoption of Electronic Health Records (EHRs) in China During the Past 10 Years: Consecutive Survey Data Analysis and Comparison of Sino-American Challenges and Experiences. Journal of Medical Internet Research. 2021 Feb;23(2):e24813. https://doi.org/10.2196/24813.

[9] Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. Journal of the American Medical Informatics Association. 2017 Nov;24(6):1142–1148. https://doi.org/10.1093/jamia/ocx080.

[10] Esdar M, Hüsers J, Weiß JP, Rauch J, Hübner U. Diffusion dynamics of electronic health records: A longitudinal observational study comparing data from hospitals in Germany and the United States. International Journal of Medical Informatics. 2019 Nov;131:103952. https://doi.org/10.1016/j.ijmedinf.2019.103952.

[11] Kim YG, Jung K, Park YT, Shin D, Cho SY, Yoon D, et al. Rate of electronic health record adoption in South Korea: A nation-wide survey. International Journal of Medical Informatics. 2017 May;101:100–107. https://doi.org/10.1016/j.ijmedinf.2017.02.009.

[12] Dreyer NA. Registries for Robust Evidence. JAMA. 2009 Aug;302(7):790. https://doi.org/10.1001/jama.2009.1092.

[13] Gehring S, Eulenfeld R. German Medical Informatics Initiative: Unlocking Data for Research and Health Care. Methods of Information in Medicine. 2018 Jul;57(S 01):e46–e49. https://doi.org/10.3414/ME18-13-0001.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[14] Semler S, Wissing F, Heyder R. German Medical Informatics Initiative: A National Approach to Integrating Health Data from Patient Care and Medical Research. Methods of Information in Medicine. 2018 Jul;57(S 01):e50–e56. https://doi.org/10.3414/ME18-03-0003.

[15] MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. Journal of the American Medical Informatics Association. 2012 Jun;19(e1):e119–e124. https://doi.org/10.1136/amiajnl-2011-000508.

[16] Zhou X, Chen S, Liu B, Zhang R, Wang Y, Li P, et al. Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. Artificial Intelligence in Medicine. 2010 Feb;48(2-3):139–152. https://doi.org/10.1016/j.artmed.2009.07.012.

[17] Pavlenko E, Strech D, Langhof H. Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. BMC Medical Informatics and Decision Making. 2020 Dec;20(1):157. https://doi.org/10.1186/s12911-020-01177-z.

[18] Khoumbati K, Themistocleous M. Integrating the IT Infrastructures in Healthcare Organisations: a Proposition of Influential Factors. Electronic Journal of Electronic Government. 2006;4(1):27–36.

[19] De Mul M, Alons P, Van Der Velde P, Konings I, Bakker J, Hazelzet J. Development of a clinical data warehouse from an intensive care clinical information system. Computer Methods and Programs in Biomedicine. 2012 Jan;105(1):22–30. https://doi.org/10.1016/j.cmpb.2010.07.002.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[20] Doutreligne M, Degremont A, Jachiet PA, Lamer A, Tannier X. Good practices for clinical data warehouse implementation: A case study in France. PLOS Digital Health. 2023 Jul;2(7):e0000298. https://doi.org/10.1371/journal.pdig.0000298.

[21] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association. 2010 Mar;17(2):124–130. https://doi.org/10.1136/jamia.2009.000893.

[22] Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association. 2010 Mar;17(2):131–135. https://doi.org/10.1136/jamia.2009.002691.

[23] Shin SY, Kim WS, Lee JH. Characteristics Desired in Clinical Data Warehouse for Biomedical Research. Healthcare Informatics Research. 2014;20(2):109. https://doi.org/10.4258/hir.2014.20.2.109.

[24] Überall MA, Müller-Schwefe GHH, Horlemann J. Efficacy and tolerability of the antispasmodic, pridinol, in patients with muscle-pain – results of primepain, a retrospective analysis of open-label real-world data provided by the German pain E-registry. Current Medical Research and Opinion. 2022 Jul;38(7):1203–1217. https://doi.org/10.1080/03007995.2022.2077579.

[25] Breuckmann F, Rassaf T, Hochadel M, Giannitsis E, Münzel T, Senges J. German chest pain unit registry: data review after the first decade of certification. Herz. 2021 Apr;46(S1):24–32. https://doi.org/10.1007/s00059-020-04912-4.

[26] (DKG) DK, GKV-Spitzenverband, der privaten Krankenversicherung (PKV) V, für das Entgeltsystem im Krankenhaus (InEK GmbH) I.: Deutsche Kodierrichtlinien Version 2024.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[27] Kolbrink B, Scheikholeslami-Sabzewari J, Borzikowsky C, Von Samson-Himmelstjerna FA, Ullmann AJ, Kunzendorf U, et al. Evolving epidemiology of pneumocystis pneumonia: Findings from a longitudinal population-based study and a retrospective multi-center study in Germany. The Lancet Regional Health - Europe. 2022 Jul;18:100400. https://doi.org/10.1016/j.lanepe.2022.100400.

[28] Lang S, Walter N, Freigang V, Neumann C, Loibl M, Alt V, et al. Increased incidence of vertebral fractures in German adults from 2009 to 2019 and the analysis of secondary diagnoses, treatment, costs, and in-hospital mortality. Scientific Reports. 2023 Apr;13(1):6984. https://doi.org/10.1038/s41598-023-31654-0.

[29] Rupp M, Walter N, Lau E, Worlicek M, Kurtz SM, Alt V. Recent trends in revision knee arthroplasty in Germany. Scientific Reports. 2021 Jul;11(1):15479. https://doi.org/10.1038/s41598-021-94988-7.

[30] Richter D, Weber R, Eyding J, Bartig D, Misselwitz B, Grau A, et al. Acute ischemic stroke care in Germany – further progress from 2016 to 2019. Neurological Research and Practice. 2021 Dec;3(1):14. https://doi.org/10.1186/s42466-021-00115-2.

[31] Walter N, Hierl K, Brochhausen C, Alt V, Rupp M. The epidemiology and direct healthcare costs of aseptic nonunions in Germany – a descriptive report. Bone & Joint Research. 2022 Aug;11(8):541–547. https://doi.org/10.1302/2046-3758.118.BJR-2021-0238.R3.

[32] Hermanns I, Kouka M, Schlattmann P, Guntinas-Lichius O. Mortality during In-Hospital Treatment for Head and Neck Cancer in Germany: A Diagnosis-Related Group-Based Nationwide Analysis, 2005–2018. Journal of Oncology. 2022 Sep;2022:1–8. https://doi.org/10.1155/2022/1387860.

medRxiv preprint doi: https://doi.org/10.1101/2024.09.13.24313333; this version posted September 18, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

[33] Heers H, Stay D, Wiesmann T, Hofmann R. Urolithiasis in Germany: Trends from the National DRG Database. Urologia Internationalis. 2022;106(6):589–595. https://doi.org/10.1159/000520372.

[34] Strumann C, Blickle PG, Von Meißner WCG, Steinhäuser J. The use of routine data from primary care practices in Germany to analyze the impact of the outbreak of SARS-CoV-2 on the utilization of primary care services for patients with type 2 diabetes mellitus. BMC Primary Care. 2022 Dec;23(1):327. https://doi.org/10.1186/s12875-022-01945-y.

[35] Guo LL, Morse KE, Aftandilian C, Steinberg E, Fries J, Posada J, et al. Characterizing the limitations of using diagnosis codes in the context of machine learning for healthcare. BMC Medical Informatics and Decision Making. 2024 Feb;24(1):51. https://doi.org/10.1186/s12911-024-02449-8.